mathematics

MDPI

Article

Heart Rate Estimation Algorithm Integrating Long and Short-Term Temporal Features

Jie Sun [1], Zhanwang Zhang [1], Jiaqi Liu [1], Lijian Zhou [1] and Songtao Hu [2,*]

[1] School of Information and Control Engineering, Qingdao University of Technology, Qingdao 266033, China; sunjie1979@qut.edu.cn (J.S.); zhangzw2310@163.com (Z.Z.); ljq202322050828@163.com (J.L.); zhoulijian@qut.edu.cn (L.Z.)

[2] School of Environmental and Municipal Engineering, Qingdao University of Technology, Qingdao 266033, China

* Correspondence: songtao-hu@qut.edu.cn

Abstract: Non-contact heart rate monitoring from facial videos utilizing remote photoplethysmography (rPPG) has gained significant traction in remote health monitoring. Given that rPPG captures the dynamic blood flow within the human body and constitutes a time-series signal characterized by periodic properties, this study introduced a three-dimensional convolutional neural network (3D CNN) designed to simultaneously address long-term periodic and short-term temporal characteristics for effective rPPG signal extraction. Firstly, differential operations are employed to preprocess video data, enhancing the face's dynamic features. Secondly, building upon the 3D CNN framework, multi-scale dilated convolutions and self-attention mechanisms were integrated to enhance the model's temporal modeling capabilities further. Finally, interpolation techniques are applied to refine the heart rate calculation methodology. The experiments conducted on the UBFC-rPPG dataset indicate that, compared with the existing optimal algorithm, the average absolute error (MAE) and the root mean square error (RMSE) realized significant enhancements of approximately 28% and 35%. Additionally, through comprehensive analyses such as cross-dataset experiments and complexity analyses, the validity and stability of the proposed algorithm in the task of heart rate estimation were manifested.

Keywords: rPPG; self-attention; 3D CNN

MSC: 68T07

check for updates

Citation: Sun, J.; Zhang, Z.; Liu, J.; Zhou, L.; Hu, S. Heart Rate Estimation Algorithm Integrating Long and Short-Term Temporal Features. *Mathematics* **2024**, *12*, 3444. https://doi.org/10.3390/math12213444

Academic Editor: Ioannis Tsoulos

Received: 15 October 2024
Revised: 1 November 2024
Accepted: 2 November 2024
Published: 4 November 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Heart rate is a crucial physiological parameter of the human body and is widely applied in medical fields and health monitoring. Traditional heart rate measurement is mainly achieved through electrocardiogram (ECG) [1] and photoplethysmograph (PPG) [2] signals. Although these two techniques are mature and accurate, they require direct skin contact. Accurately obtaining ECG and PPG signals from patients with skin conditions and newborn infants poses significant challenges, making real-time monitoring difficult. The advancement and widespread adoption of artificial intelligence technology and telemedicine indicate that non-contact, real-time heart rate measurement methods are poised to become a significant trend in future developments. Non-contact physiological parameter measurement methods based on rPPG technology have developed rapidly in recent years. Researchers have successfully applied rPPG signals to measure multiple physiological parameters such as heart rate, heart rate variability, respiratory rate, and blood oxygen level [3–7]. However, regardless of which physiological parameter is measured, extracting a robust and accurate rPPG signal from the facial video is necessary. However, the motion artifacts and light intensity usually influence the facial video collection. Therefore, extracting stable and accurate rPPG signals constitutes a challenging task.

In 2008, Verkruysse et al. [8] were the first to obtain the periodic variations in the skin surface color by detecting consecutive facial videos and successfully extracting the remote PPG signal, namely the rPPG signal. This approach has established a novel direction for non-contact physiological signal detection and has become the foundation for subsequent research. The rPPG signal extraction algorithm has progressively evolved with ongoing technological advancements, leading to various implementation pathways. Currently, rPPG signal extraction algorithms are mainly classified into traditional and deep learning methods. Among the traditional methods, there are primarily blind source separation (BSS) algorithms (independent component analysis (ICA) [3], principal component analysis (PCA) [9]) and model-based methods (chrominance-based method (CHROM) [10], plane orthogonal to the skin (POS) [11]). BSS depends on the assumption that the source signals are independent or uncorrelated. These methods have a certain robustness to the mixture of noise and signals. However, its performance is unsatisfactory when dealing with complex nonlinear signals (such as optical property changes caused by skin pigmentation changes). Meanwhile, the model-based methods need more flexibility in response to environmental changes (such as changes in illumination) or individual differences of the subjects (such as different skin pigment concentrations, among others) and thus need help coping with these changes.

In contrast to the previously mentioned traditional methods that depend on prior knowledge for support, deep learning, which is fundamentally data-driven, demonstrates greater robustness against the diverse noise in facial data. Špetlík et al. [12] pioneered an end-to-end heart rate prediction model in which a 2D CNN served as the signal extractor, and the extracted rPPG signal was input into an HR estimator to output the predicted heart rate. Chen and McDuff [5] proposed the DeepPhys model with a VGG architecture incorporating motion representation and an attention mechanism, demonstrating greater robustness against illumination and motion artifacts. Because 2D CNN is deficient in the capacity to extract temporal information, and temporal information significantly impacts the extraction of rPPG signals, researchers have progressively inclined their research emphasis towards the extraction of temporal information. Yu et al. [13] adopted a 3D CNN network model in the rPPG signal task, which added temporal dimension modeling compared with 2D CNN, yielding a notable performance improvement. Subsequently, refs [6,14,15] all based their models on 3D CNN and enhanced the representational capabilities in corresponding dimensions by integrating attention mechanisms. Researchers have also explored the combination of CNN and RNN to better acquire temporal features from facial video data. Huang et al. [16] proposed a novel approach that combined 3D CNN with LSTM, enabling 3D CNN to extract spatial features and short-term temporal information from each video frame and LSTM to capture long-term temporal information from consecutive frames. Wang [17] employed a two-stream network for feature extraction and signal extraction. Based on 2D CNN, the feature extraction stream was utilized to extract synchronous spatial features from the spatial-temporal map. The signal extraction stream consisted of a combination of 2D CNN and LSTM, where 2D CNN was applied to extract the initial rPPG signal, and two layers of LSTM were used for further refinement of the initial signal. The outputs of the two streams were combined and served as the input for the HR estimator. While LSTM is adept at capturing certain long-term dependencies, it may encounter challenges such as gradient vanishing or explosion when processing extremely lengthy time series, limiting its ability to capture long-term temporal characteristics effectively. In contrast, the Transformer [18] model leverages a self-attention mechanism to acquire global contextual information more efficiently, thereby providing significant advantages in extracting rPPG signals.

Models leveraging the Vision Transformer [19], including the Physformer model introduced by Yu et al. [20] and the EfficientPhys model developed by Liu et al. [21], have demonstrated commendable performance in this study. However, within the current landscape of rPPG datasets, a large-scale dataset that can fully exploit the performance potential of the Transformer architecture has yet to materialize. In response to this challenge,

Yu et al. employed model pre-training and data augmentation techniques. Considering the unique characteristics of rPPG data and the constraints posed by available data approaches that integrate convolutional neural networks (CNNs) with self-attention mechanisms exhibit significant advantages.

This study presented an adaptable rPPG signal extraction method built upon an encoder–decoder structure with a 3D CNN backbone network. The addition of a long short-term temporal self-attention module (LST-SAM) to the backbone network enhances its adaptability. This module uses parallel multi-scale dilated convolutions for short-term temporal dependency modeling. The multi-scale short-term temporal features are first fused. Then, a self-attention mechanism is applied to assign appropriate attention weights to long-term temporal dependencies. Finally, skip connections complement the backbone network's spatial-temporal encoding features to achieve long short-term temporal feature modeling. The adaptability of this method makes it a promising solution for a variety of rPPG tasks.

2. Long Short-Term Attention rPPG Network

Figure 1 presents the overall framework of the heart rate estimation algorithm based on the fusion of long-term and short-term temporal features proposed in this study. Firstly, the face region of the face video undergoes preprocessing such as detection, cropping, and differencing. Secondly, the preprocessed data are input into the Long Short-Term Attention rPPG Network (LSTA-rPPGNet) to extract rPPG signals. Subsequently, signal filtering and the Fast Fourier Transform (FFT) are used to calculate the power spectral density (PSD). The Akima interpolation method then interpolates the PSD in the frequency domain to determine the power value representing the heart rate. Finally, the heart rate is computed.

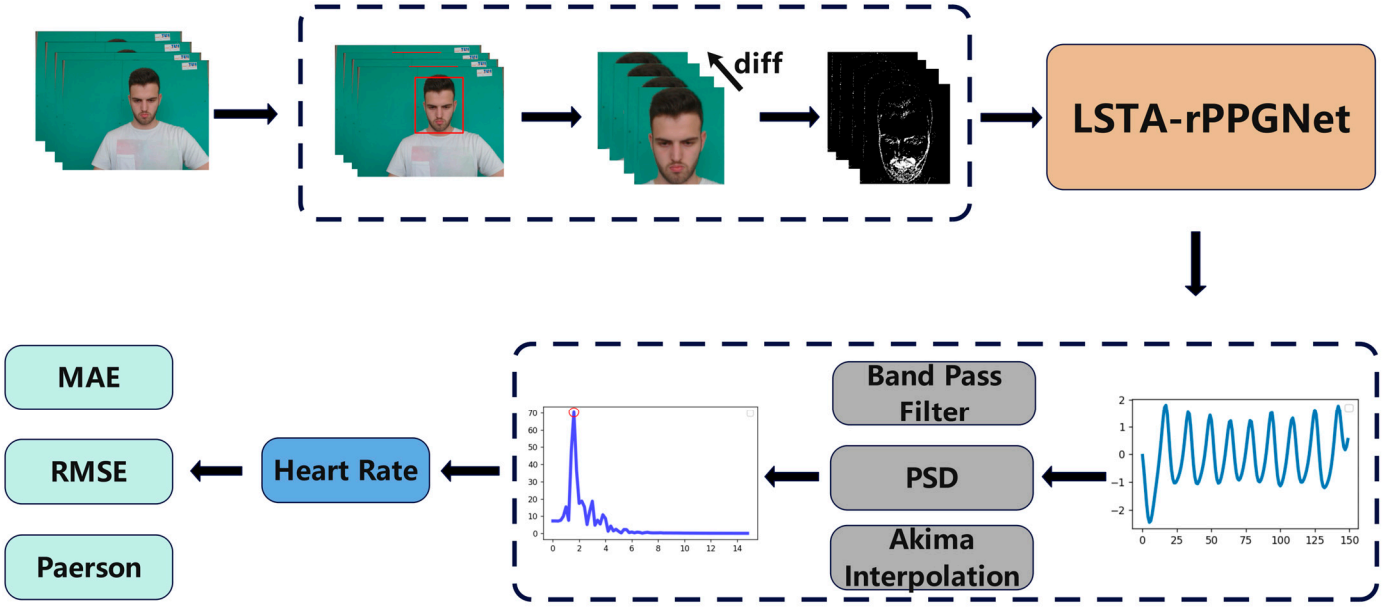


Figure 1. The overall framework of the algorithm.

2.1. LSTA-rPPGNet

Three-dimensional convolutional neural networks (3D CNNs) can effectively capture temporal dynamics and spatial features in video data by performing convolutional operations across three dimensions (width, height, and time). This ability is crucial for understanding dynamic changes in videos, such as subtle changes in facial expressions. In contrast, 2D convolutional neural networks (2D CNNs) typically only capture spatial features in static images and cannot fully utilize information in the time dimension. Consequently, this study employs a model based on 3D CNN to achieve a more precise understanding and analysis of dynamic changes present in video content.

Figure 2 depicts the rPPG signal extraction network's overall framework proposed in this paper, with the backbone network adopting an encoder–decoder architecture. The encoder progressively extracts temporal and spatial features from video frames through a sequence of 3D convolutional blocks (3D convolution, batch normalization, and ReLU activation function). Moreover, multiple 3D max-pooling operations are applied to the feature map for down-sampling, enabling the extraction of multi-scale features while preserving essential temporal and spatial information. During the feature encoding, a long short-term temporal self-attention module is added through skip connections to enhance the model's ability to extract long-term sequential features. The feature map's spatial and temporal resolutions are reduced during encoding. In the decoder part, the network employs deconvolution layers to gradually restore the temporal resolution and restore the temporal information to the original frame rate through the up-sampling operations of the feature map. Eventually, an rPPG signal matching the temporal length of the input video is generated using adaptive average pooling and 1×1 convolutional layers.

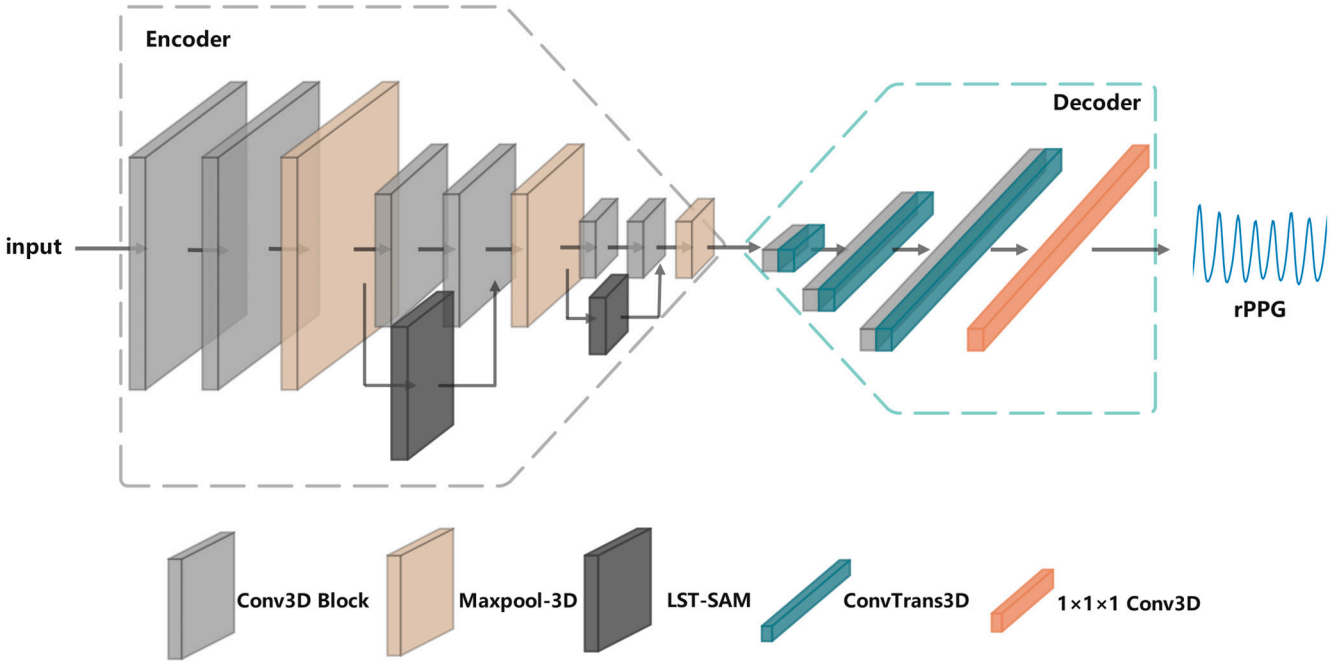


Figure 2. The model diagram of Long-Short Term Attention rPPG Network (LSTA-rPPGNet), where Conv3D Block represents a sequence of 3D convolutional blocks (3D convolution, batch normalization, and ReLU activation function). Maxpool-3D represents 3D max pooling, LST-SAM represents Long-Short Term Self-Attention Module, and ConvTrans3D represents 3D transposed convolution.

Table 1 provides a detailed description of the network structure. The name indicates the operation within each layer structure. Conv3D represents the 3D convolution structure, and 3D BatchNorm and ReLu operations follow each 3D convolution. ConvTrans3D represents the three-dimensional transposed convolution, and MaxPool represents the 3D max pooling operation. Except Conv3D-10, which uses a stride of (1,1,1) and zero padding, all other 3D convolution operations are performed with a stride of (1,1,1) and padding of (1,1,1). In contrast, the 3D transposed convolution operation is carried out with a stride of (2,1,1) and padding of (1,0,0). Kernel represents the size of the convolution kernel, where $16 \times 3 \times 3 \times 3$ indicates 16 convolution kernels of size $3 \times 3 \times 3$. Output size represents the size of the features output by each layer operation, namely the number of channels, time, height, and width.

Table 1. Network structure.

Name	Kernel	Output Size
Input	None	$3 \times T \times 72 \times 72$
Conv3D-1	$16 \times 3 \times 3 \times 3$	$16 \times T \times 72 \times 72$
Conv3D-2	$16 \times 3 \times 3 \times 3$	$16 \times T \times 72 \times 72$
MaxPool-1	$2 \times 2 \times 2$	$16 \times T/2 \times 36 \times 36$
Conv3D-3	$32 \times 3 \times 3 \times 3$	$32 \times T/2 \times 36 \times 36$
Conv3D-4	$32 \times 3 \times 3 \times 3$	$32 \times T/2 \times 36 \times 36$
MaxPool-2	$2 \times 2 \times 2$	$32 \times T/4 \times 18 \times 18$
Conv3D-5	$64 \times 3 \times 3 \times 3$	$64 \times T/4 \times 18 \times 18$
Conv3D-6	$64 \times 3 \times 3 \times 3$	$64 \times T/4 \times 18 \times 18$
MaxPool-3	$2 \times 2 \times 2$	$64 \times T/8 \times 9 \times 9$
Conv3D-7	$64 \times 3 \times 3 \times 3$	$64 \times T/8 \times 9 \times 9$
ConvTrans3D	$64 \times 4 \times 1 \times 1$	$64 \times T/4 \times 9 \times 9$
Conv3D-8	$64 \times 3 \times 3 \times 3$	$64 \times T/4 \times 9 \times 9$
ConvTrans3D	$64 \times 4 \times 1 \times 1$	$64 \times T/2 \times 9 \times 9$
Conv3D-9	$64 \times 3 \times 3 \times 3$	$64 \times T/2 \times 9 \times 9$
ConvTrans3D	$64 \times 4 \times 1 \times 1$	$64 \times T \times 9 \times 9$
Conv3D-9	$64 \times 3 \times 3 \times 3$	$64 \times T \times 9 \times 9$
MaxPool-4	$1 \times 9 \times 9$	$64 \times T \times 1 \times 1$
Conv3D-10	$1 \times 1 \times 1 \times 1$	$1 \times T \times 1 \times 1$

2.2. Long Short-Term Self-Attention Module (LST-SAM)

In extracting rPPG signals, the characteristics associated with varying time durations are critically important for analytical outcomes. Given that rPPG signals are susceptible to noise and interference from factors such as head movement and lighting variations, incorporating convolutional kernels of different temporal lengths can effectively mitigate these short-term disturbances, thereby enabling the model to focus on long-term features of physiological relevance. Specifically, shorter time windows are adept at capturing rapid signal fluctuations, such as instantaneous changes in heart rate. In comparison, longer time windows facilitate the identification of more stable physiological trends (e.g., periodic variations in heartbeats). Consequently, based on 3D CNN, we integrated the long short-term temporal self-attention module (LST-SAM) to enable the model to extract features across varying temporal scales and enhance its adaptability to complex signals.

As shown in Figure 3, this module encompasses a Short-Term Temporal Feature Extractor (STE) and a Long-Term Temporal Feature Extractor (LTE). The STE module is composed of dilated convolutions with varying dilation factors in the temporal dimension. This module effectively models features across various temporal spans in the short term by employing dilated convolutions with different dilation factors. Smaller dilation factors are more adept at capturing short-term information, while larger dilation factors facilitate the extraction of features over extended time ranges. This multi-scale processing strategy allows for the effective extraction of subtle variations and long-term trends within rPPG signals. The core mechanism of the LTE module is self-attention, which follows the STE module and employs a self-attention mechanism to assign distinct weights to features at different time steps, thereby enabling global weighting in the temporal dimension. This process emphasizes features with long-term dependencies within rPPG signals while disregarding transient noise or irrelevant information, thus facilitating comprehensive modeling of the entire time series data. The STE module provides temporally diverse feature scales for the long-term attention module, allowing it to perform global weighting based on richer temporal information and ensure that short-term details and long-term trends in rPPG signal patterns receive appropriate attention.

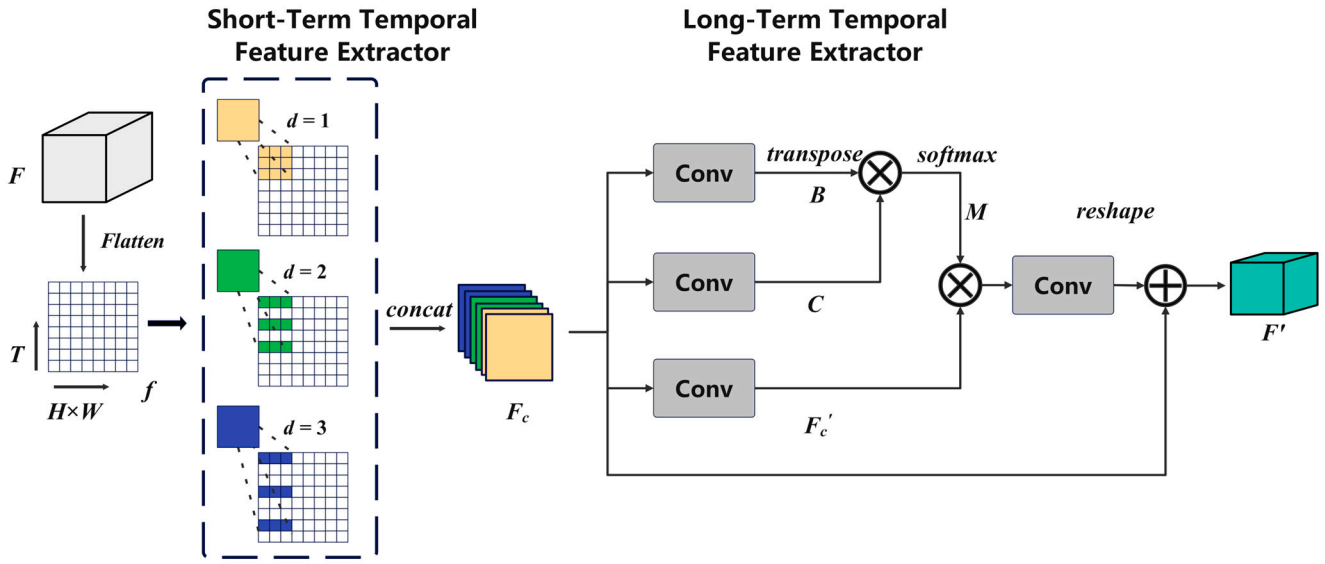


Figure 3. Long Short-Term Self-Attention Module (LST-SAM).

2.2.1. Short-Term Temporal Feature Extractor

In conventional CNNs, the convolution operation samples on consecutive pixels or feature maps, which results in a relatively limited receptive field size. Dilated convolution [22] enables a larger receptive field through interval sampling, allowing the model to perceive broader contextual information. This is highly beneficial for semantic segmentation or time series analysis tasks. In this research, parallel dilated convolution along the temporal dimension was employed. By utilizing different dilation rates, temporal features of different scales could be extracted concurrently. This multi-scale feature extraction capability enables the model to better capture the details and variations in the data. During the extraction process of rPPG signals, temporal information's continuity and integrity are paramount. Multi-scale dilated convolution reduces the likelihood of temporal information loss by retaining more input information during the convolution process, thereby enhancing the model's performance.

As shown in the Short-Term Temporal Feature Extractor in Figure 3, following the spatial dimension flattening of feature $F(C, T, H, W)$, a 2D representation $f(C, T, H \times W)$ is derived. Subsequently, f is subjected to dilation convolution with a convolution kernel W_d of size 3×3, where d is the dilation coefficient. The convolution kernel is exclusively expanded in the temporal dimension to achieve scale-invariant temporal features while maintaining spatial information. Short-term temporal features of varying time lengths can be extracted using varying degrees of dilated convolution and receptive fields of different sizes. Subsequently, the features obtained from these three dilation convolutions are concatenated. This process is aimed at obtaining F_c, which enables the acquisition of multi-scale temporal features. The above steps can be summarized as follows:

$$f = \phi(F), \tag{1}$$

$$F_c = concat(f * W_d) \;\; d = 1, 2, 3, \tag{2}$$

in which $\phi(\cdot)$ represents the flattening process of the spatial dimension, and $concat(\cdot)$ represents the feature concatenation in the channel dimension.

2.2.2. Long-Term Temporal Feature Extractor

Since the self-attention mechanism was introduced, its exceptional performance in Transformer models has clearly demonstrated its robust capability for global feature modeling. Due to its superior ability to model long-range spatial dependencies in computer vision, the self-attention mechanism has been widely employed in tasks such as image

segmentation and action recognition. This study aimed to extend the advantages of the self-attention mechanism to global temporal feature extraction and investigate its potential applications in long-term dependency modeling.

This is shown in Figure 3 of the Long-Term Temporal Feature Extractor (LTE).In this module, we incorporated the non-local mechanism [23], which fundamentally operates as a self-attention mechanism. The fundamental concept of non-local is to compute an attention mask M of size T $\times$ T to store the contextual relationships among all frame features. After obtaining the short-term temporal features F_c concatenated by dilation convolution, three 1×1 convolutional layers are initially applied to generate the feature maps B, C, and F_c'. Then, the feature map B is transposed and subjected to matrix multiplication with C, and a T $\times$ T-sized long-term attention mask is obtained using the softmax function. Finally, the long-term temporal attention weighting of F_c' is achieved through matrix multiplication with the attention mask M, and a skip connection is made with F_c. The fused features are reshaped into the same shape as F_i:

$$M = \sigma(\theta(B) \times C), \tag{3}$$

$$F_i' = \delta(W * (F_c' \times M) + F_c), \tag{4}$$

in which $\theta(\cdot)$ represents matrix transposition, $\sigma(\cdot)$ represents the softmax operation, and $\delta(\cdot)$ represents the reshape operation.

2.3. Loss Function

In the domain of heart rate estimation, the primary objective is to extract an rPPG signal that closely aligns with the shape of the authentic PPG signal. Loss functions such as mean absolute error (MAE) and mean squared error (MSE) emphasize the amplitude discrepancies between predicted values and actual measurements. However, in rPPG signals, amplitude variations may occur due to external factors such as lighting conditions and skin tone differences. The negative Pearson correlation coefficient is independent of absolute signal amplitudes and instead emphasizes the correlation between signals, thus providing greater robustness against these amplitude fluctuations. Consequently, this study employs the negative Pearson correlation coefficient as its loss function, defined by the following formula:

$$Loss = 1 - \frac{\sum_{i=1}^{T}(x(i) - \overline{x})(y(i) - \overline{y})}{\sqrt{\sum_{i=1}^{T}(x(i) - \overline{x})^2}\sqrt{\sum_{i=1}^{T}(y(i) - \overline{y})^2}}, \tag{5}$$

In this equation, $x(i)$ denotes the predicted rPPG signal at time i, $y(i)$ denotes the actual rPPG signal at time i, and T represents the length of the rPPG signal. Both $\overline{x}$ and $\overline{y}$ signify the mean values of their respective signals. The Pearson correlation coefficient, which ranges from -1 to +1, quantifies the similarity between these two signals. A correlation value of -1 indicates a perfect negative correlation between the two signals, while 0 signifies no linear correlation. Conversely, a value of +1 indicates a perfect positive correlation. This study aimed to predict an rPPG signal strongly correlated with the ground truth PPG signal, with an optimal goal of achieving a correlation coefficient as close to 1 as possible.

3. Experiment and Results Analysis

To assess the advantages of the proposed algorithm, we conducted comparative experiments against other algorithms using two benchmark datasets. Additionally, ablation studies were performed to evaluate the effectiveness of the proposed LST-SAM module. All models discussed in this section were trained on PyTorch utilizing NVIDIA A800 GPUs. This study employed the Adam optimizer with a learning rate set at 5×10^{-4}. The performance of the algorithm was assessed using three metrics: mean absolute error (MAE), root mean square error (RMSE), and Pearson correlation coefficient (r)

3.1. Dataset

UBFC-rPPG [24] is a dataset for rPPG analysis, encompassing 42 videos of 42 subjects (31 males and 11 females), along with the corresponding ground truth rPPG values and heart rates. Each video has an approximate duration of 2 minutes and was acquired at a rate of 30 frames per second via a low-cost webcam (Logitech C920 HD Pro) with a resolution of 640 × 480 in uncompressed 8-bit RGB format. The PPG signals and HR were acquired using a CMS50E transmissive pulse oximeter as the ground truth values for the videos. The videos were recorded indoors with different natural light and illuminations. The subjects were seated approximately one meter away from the camera with their faces fully visible, and they were required to participate in a time-sensitive math game, aiming to increase their heart rates and thereby simulate a natural human-computer interaction scenario.

PURE [25] dataset comprises ten subjects (eight males and two females). Each subject has six distinct activities (steady state, speaking, slow head rotation, fast head rotation, small-scale head rotation, and medium-scale head rotation). There are 60 videos, each approximately 1 min in duration. All the videos are captured by an industrial camera at 30 fps, with a spatial resolution of 640 × 480 pixels. The pulox CMS50E finger pulse oximeter acquires the ground truth PPG signal at a sampling rate of 60 Hz.

3.2. Video Data Preprocessing

This study conducted comparative experiments on video sample length, facial region resolution size, and data augmentation based on differential operations to verify the necessity of preprocessing strategies. Through these experiments, we were able to thoroughly assess the impact of different parameter settings on data quality and model performance. These provided a solid foundation for finalizing the preprocessing strategy and ensured the reliability and validity of the experimental results.

3.2.1. Video Sample Length Comparison Experiment

The extraction of rPPG signals from facial videos is significantly influenced by the length of the video sample. Shorter segments may inadequately capture sufficient heartbeat cycles, resulting in unstable or incomplete rPPG signals. Conversely, longer videos can offer more heartbeat cycles to enhance signal stability and accuracy, but they also increase noise and computational complexity and time. This study conducted a comparative analysis involving 72, 128, 180, and 240 frames, with the facial region size standardized to 72 × 72 pixels to ascertain an optimal video length. Figure 4 illustrates the results.

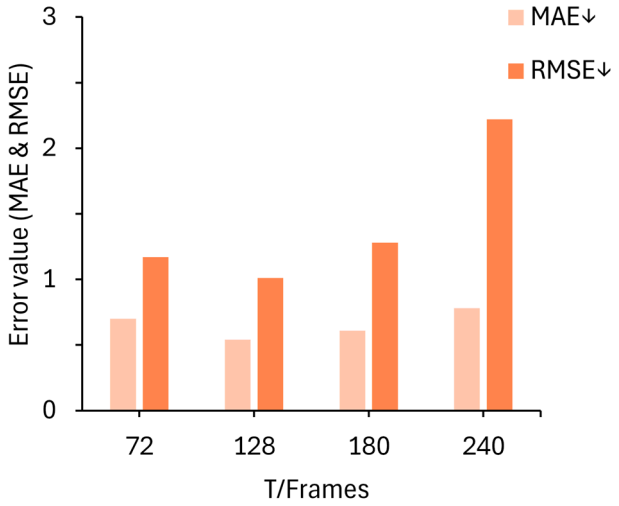


Figure 4. Video length comparison experiment. (The downward arrows next to MAE and RMSE in the legend indicate that lower values are better, meaning that smaller error values represent improved performance).

As illustrated in Figure 4, when the sample length is 128 frames, the MAE and RMSE attain their minimum values, resulting in optimal experimental outcomes. This phenomenon can be attributed to the normal human heart rate, which typically ranges from 60 to 120 beats per minute (i.e., approximately 1 to 1.67 beats per second). Considering a video frame rate of 30 fps, the duration corresponding to 128 frames is roughly 4.27 s, allowing for coverage of approximately 4 to 7 heartbeat cycles. The selection of a duration of 128 frames facilitates the extraction of complete and stable rPPG signals while minimizing excessive motion artifacts and mitigating long-term signal drift effects. This specific length effectively balances signal quality with computational complexity. In contrast, shorter video lengths (e.g., 72 frames) may provide inadequate coverage of heartbeat cycles, leading to unstable signal extraction; conversely, longer videos (e.g., 180 or 240 frames), while offering more information, significantly escalate computational complexity and time costs and may also result in capturing irrelevant noise or motion artifacts. Therefore, this study adopts a preprocessing strategy establishing each video's sample length at 128 frames.

3.2.2. Resolution Comparison Experiment

The resolution size of the facial area significantly impacts the experimental outcomes in rPPG signal extraction. High resolution enhances the signal's precision and quality; however, it also increases computational complexity and is more susceptible to lighting conditions and motion artifacts. Conversely, lower resolution may improve processing speed and robustness but can result in a loss of critical signal details. Following determining video length, a comparative experiment was conducted on facial area size while maintaining a fixed video length of 128 frames. As illustrated in Figure 5, the face was initially located using four specific positions among the 68 landmarks, after which the facial area was enlarged by 1.5 times and cropped, as shown in Figure 5b. Subsequently, resolutions were adjusted to sizes of 48 × 48, 72 × 72, 128 × 128, and 156 × 156 respectively, as presented in Figure 5c. The results from these experiments are depicted in Figure 6.

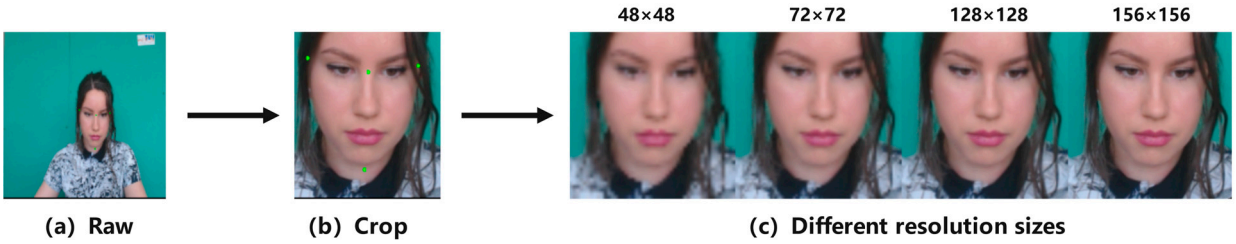


Figure 5. Facial region preprocessing.

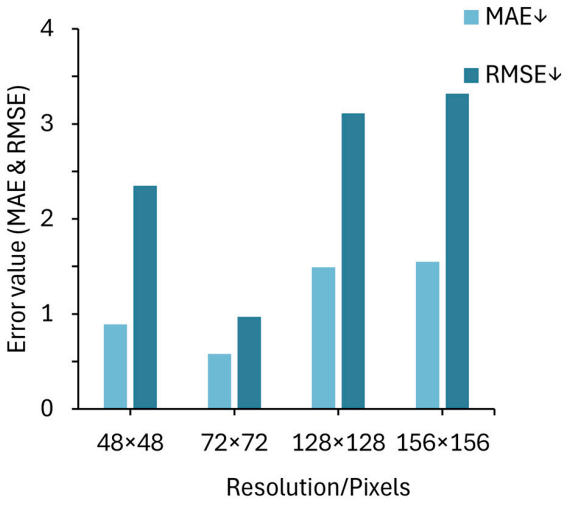


Figure 6. Comparison of resolution size. (The downward arrows next to MAE and RMSE in the legend indicate that lower values are better, meaning that smaller error values represent improved performance).

As can be observed from Figure 6, the facial area resolution of 72 × 72 pixels yields the most favorable results. In larger images (e.g., 128 × 128 pixels), despite the higher resolution, these additional pixels may not provide more valuable signal information; instead, they could increase data redundancy, rendering the rPPG signal more susceptible to noise during processing. Furthermore, rPPG signals predominantly depend on overall color variations of the face instead of precise pixel-level details. Consequently, a lower resolution (such as 72 × 72 pixels) is likely sufficient to capture these changes without introducing extraneous detail noise. A resolution of 48 × 48 pixels may be low, leading to significant image distortion and substantial loss of spatial information, thereby hindering feature extraction. Therefore, this study adopts a preprocessing strategy that sets the facial area resolution at 72 × 72 pixels.

3.2.3. Data Augmentation (Overlapping, Difference)

A video segment was created to augment the data samples by moving every 30 frames, with an overlap of 98 frames between adjacent segments. An inter-frame difference operation was performed on the original data to capture the minute color variations in the face between consecutive video frames and amplify these subtle changes. We conducted a comparative experiment to validate the efficacy of the inter-frame difference operation under the condition of fixed sample length and size. Table 2 presents the results.

Table 2. Differential operation comparison experiment (✓ represents differential, while ✗ represents non-differential).

Difference	MAE	RMSE	Pearson
✗	0.76	1.68	0.98
✓	0.58	0.97	0.99

The results in Table 2 show that the facial video data after differential processing exhibited a marked improvement compared with that without difference processing. The inter-frame difference operation can effectively mitigate the interferences caused by lighting variations and facial movements, enhance the weak signal features related to blood flow, and enhance the signal-to-noise ratio and extraction precision.

3.3. Optimization of Heart Rate Algorithm

In numerous prior studies, researchers typically began by applying bandpass filtering to the estimated rPPG signal to eliminate noise and interference before calculating the maximum power spectral density of the rPPG signal. The corresponding frequency identified was considered the dominant frequency indicative of heart rate. However, this approach has a significant limitation: the frequency resolution of the Fast Fourier Transform (FFT) is constrained by the length of the signal, resulting in relatively low-frequency resolution that may hinder the precise identification of frequency components.

In light of this limitation, this study implemented a 0.75–2.5 Hz bandpass filter on the PPG signal and applied symmetric zero-padding at both ends of the signal. Following spectrum calculation for the PPG signal, Akima [26] interpolation was performed. Among these operations, symmetric zero-padding effectively enhances frequency resolution, improving accuracy in identifying frequency components—a critical factor in high-precision scenarios such as heart rate detection. Akima interpolation is a local method that produces smooth transitions between data points, effectively addressing minor signal amplitude variations while avoiding oscillation artifacts. Following the computation of the PSD, Akima interpolation can be applied to achieve smooth interpolation along the frequency axis, thereby enhancing the clarity of peaks within the signal that represent heart rate frequency components. This improvement facilitates more accurate localization of peak frequencies in the spectrum, ultimately increasing the precision of heart rate calculations. Since rPPG signals are often affected by ambient light and motion noise, irregular fluctuations can

occur in the spectrum. Akima interpolation effectively handles boundaries and discontinuities, reducing the impact of anomalous data points and irregularities. Consequently, PSD curves processed through Akima interpolation provide a more accurate representation of the actual spectral characteristics of signals, reducing frequency deviations and errors associated with noise. The results are summarized in Table 3:

Table 3. Comparison of algorithm optimization experiments (✓ represents Akima interpolation, while ✗ represents non- interpolation).

Optimization	MAE	RMSE	Pearson
✗	0.63	2.11	0.98
✓	0.58	0.97	0.99

Table 3 above shows that, after Akima interpolation, the heart rate's RMSE error is significantly reduced compared with the case without interpolation. The heart rate peaks can be positioned more precisely on the frequency axis through symmetric padding and spectrum interpolation, thereby reducing extreme errors in heart rate estimation. Such interpolation can effectively correct more significant errors, reducing extreme errors and decreasing the RMSE.

3.4. Comparison Experiments with Different Methods

We initially conducted independent training and testing on the UBFC-rPPG to validate the proposed algorithm's performance. We partitioned 42 video sets into training and testing sets at a ratio of 7:3. Concurrently, we merely adopted the overlapping video segment cropping approach described in the preprocessing mentioned above for the training data. For the testing data, the created video segments don't overlap.

In our implementation, we utilized the rPPG open-source toolkit [27] to implement two traditional methods, POS [11] and CHROM [10], and deep learning-based methods such as PhysNet [13], DeepPhys [5], and EfficientPhys [21]. PhysNet employed a 3D CNN encoder and decoder structure similar to the baseline architecture in our study, with the same preprocessing method as in this study. DeepPhys adopted a combination of Motion Model and Appearance Model branches to enhance the feature modeling capability of the algorithm, maintaining the original preprocessing method. EfficientPhys proposed two approaches, including the CNN version and the Transformer version, and it is precisely the CNN version that was replicated in this study. This method not only performed differential processing but also conducted batch norm processing on the data. The network model employed temporal shift modules and self-attention mechanisms to model spatial-temporal features.

As shown in the results of the comparative experiment in Table 4, compared with the two traditional methods, POS and CHROM, the deep learning-based methods have lower MAE and RMSE, indicating that the deep learning-based methods have higher accuracy and stability. Among them, DeepPhy significantly enhanced the dynamic feature modeling capability of the algorithm through the Motion Model branch, and both MAE and RMSE reduced considerably. EfficientPhys amplified the dynamic differences of facial video data through differential and batch norm processing. Combining the time shift module and the self-attention mechanism exhibits better performance. This model emphasizes temporal features more than DeepPhys, highlighting the significance of temporal features for the rPPG signal extraction task. PhysNet employs a 3D CNN encoder–decoder structure combined with the cropping, differencing, and other preprocessing approaches adopted in this study; this network achieves a superior error performance compared to DeepPhy and EfficientPhys. The above comparative studies demonstrate that 3D CNNs have a distinct advantage over 2D CNNs in extracting rPPG from facial videos and can exploit the implicit spatiotemporal information in the video data to the greatest extent possible. Compared with all other methods, the method proposed in this study exhibits superior performance in all performance indicators, with MAE at 0.58, RMSE at 0.97, and r at 0.99.

Table 4. Experimental results on the UBFC-rPPG dataset.

Method	MAE	RMSE	Pearson
POS	2.44	6.61	0.94
CHROM	2.67	4.61	0.96
DeepPhys	0.92	2.59	0.97
EfficientPhys	0.87	2.38	0.98
PhysNet	0.81	1.54	0.98
Ours	0.58	0.97	0.99

Table 5 illustrates the complexity analysis of PhysNet, DeepPhys, EfficientPhys, and our proposed method. Specifically, our method comprises 1.05 M parameters and 26.32 G FLOPs, which presents a slightly higher computational complexity than PhysNet. However, it requires less computation than EfficientPhys and DeepPhys while significantly improving accuracy over PhysNet.

Table 5. Complexity comparison.

Method	Params	FLOPs
PhysNet	0.77 M	17.70 G
DeepPhys	2.23 M	163.54 G
EfficientPhys	2.16 M	82.66 G
Ours	1.05 M	26.32 G

Although PhysNet exhibits optimal performance in terms of FLOPs, its relatively low parameter count may constrain its learning capacity. Conversely, although DeepPhys has the highest parameter count, it also experiences a significant increase in FLOPs—suggesting that more parameters do not always lead to better performance and may result in overfitting and inefficient resource utilization. In contrast, our method attains superior accuracy while maintaining comparatively lower computational complexity, thereby validating its effectiveness and practicality.

3.5. Cross-Dataset Experiment

Cross-dataset validation enables a more comprehensive evaluation of the model's generalization ability, robustness, and credibility, particularly in cases involving diverse application scenarios, and holds significant practical significance. Hence, this study conducted a cross-dataset experiment of training on the PURE dataset and testing on the UBFC-rPPG dataset to verify the generalization ability of the proposed method. The strategies adopted for the training and testing data were the same as those mentioned above, except that due to the more significant amount of test data, considering the entire video as one heart rate sample was employed when calculating the heart rate error. Table 6 presents the experimental results.

Table 6. Cross-dataset experimental results (PURE training, UBFC-rPPG testing).

Method	MAE	RMSE	Pearson
PhysNet	2.30	5.97	0.94
DeepPhys	1.51	3.82	0.97
EfficientPhys	2.15	3.43	0.97
Ours	1.45	3.08	0.98

From Table 6, the performance of various methods shows varying degrees of increase in MAE and RMSE when evaluated in cross-dataset experiments compared with their performance on a single dataset. Nevertheless, the proposed method demonstrates superior performance across several error metrics relative to other deep learning approaches,

achieving an MAE of 1.45, an RMSE of 3.08, and a Pearson correlation coefficient of 0.98. This result indicates that the features extracted by the proposed algorithm exhibit enhanced robustness and can capture universal characteristics related to rPPG signals applicable across different datasets. Furthermore, this suggests that the algorithm possesses a heightened capacity for feature extraction, enabling it to identify critical information within the data effectively.

3.6. Visualization Analysis

Figure 7 shows the overall performance of our heart rate prediction model. Each point on the chart corresponds to a single independent prediction result. Ideally, all points should cluster along the line y = x, meaning that the predicted value is the same as the actual value. The graph clearly illustrates that most predicted points are tightly clustered around the ideal line, indicating a high level of accuracy in our model. This dense distribution underscores the reliability and stability of the model in heart rate prediction tasks, confirming that our algorithm effectively extracts heart rate information from physiological signals and facilitates precise predictions.

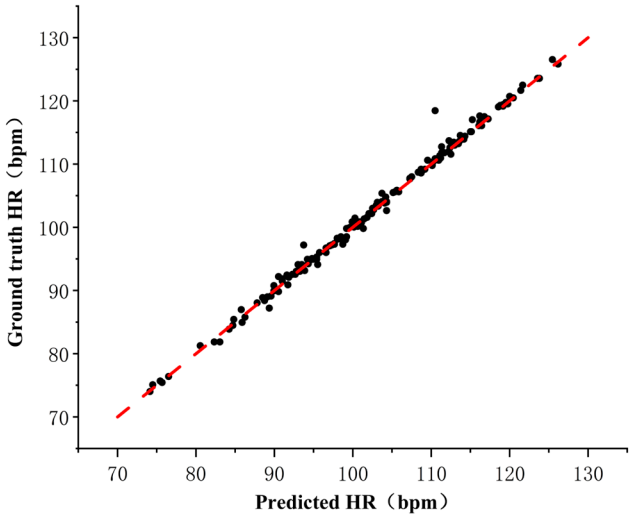


Figure 7. Comparison of predicted HRs with ground-truth HRs for testing data. The *x*-axis represents the predicted heart rate (Predicted HR (bpm)), and the *y*-axis represents the actual heart rate (Ground truth HR (bpm)).

In Figure 8a, the predicted rPPG signal is aligned with the ground truth signal in the temporal dimension, especially regarding overall trends and waveform shapes. The expected signal can roughly reflect the ground truth signal 's changing pattern. The central peak frequency positions are consistent in the frequency spectrum. The primary frequency component's degree of matching between the predicted and ground truth signal is high. Figure 8b shows the ground truth signal with apparent fluctuations in the time domain waveform. They indicate that the rPPG signal is affected by various factors in actual applications and may contain certain noise and irregular changes. However, the predicted signal in the overall trend and waveform is highly consistent with the ground truth signal, even maintaining consistency in the fluctuating areas. The central peak frequencies of the ground truth signal and predicted signals are aligned in the frequency spectrum. This result shows that our model still can stably extract the main physiological signals in complex environments with high accuracy and robustness.

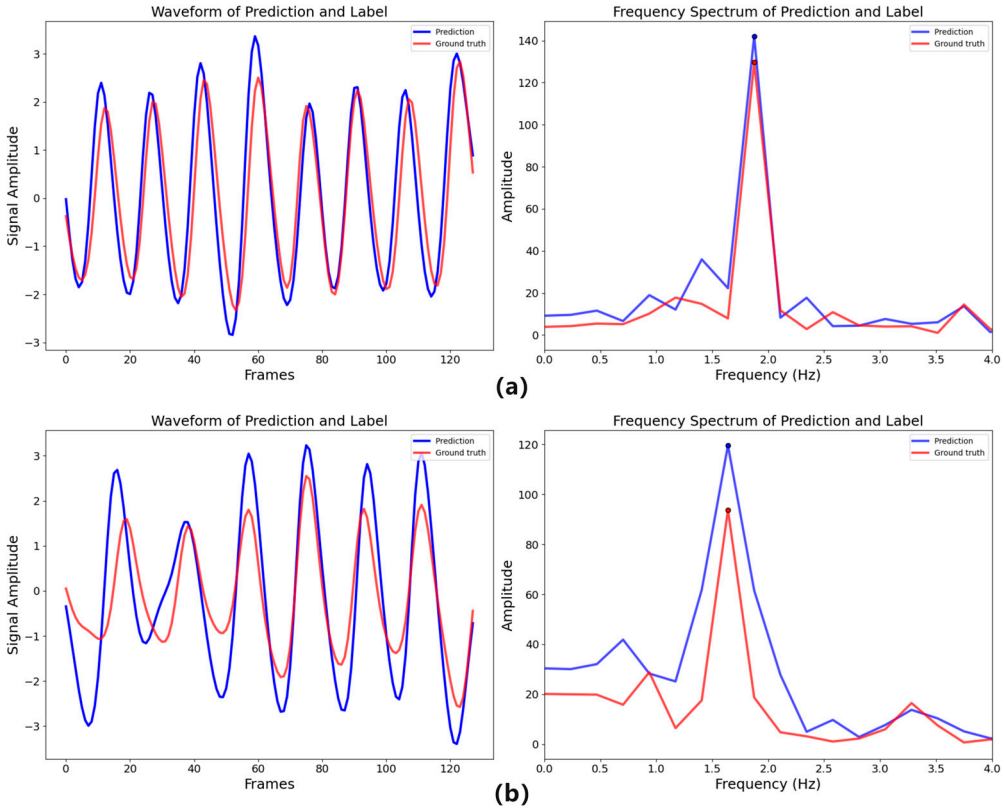


Figure 8. Comparison of predicted rPPG signal (blue) and ground truth signal (red): (**a**) obtained from subject 4; (**b**) obtained from subject 10.

3.7. Ablation Experiment

This study conducted a series of ablation experiments to verify the contribution of the long-term and short-term temporal attention modules to the final result. The experimental results are shown in Table 7.

Table 7. Results of the ablation experiment.

STE	LTE	MAE	RMSE
✗	✗	0.95	1.65
✓	✗	0.90	1.45
✗	✓	0.75	1.35
✓	✓	**0.58**	**0.97**

In the Table 7, the best-performing values are highlighted in bold to indicate the lowest error values, representing superior performance. The symbols ✓ and ✗ indicate whether the module was applied or not, respectively. Compared with the baseline architecture, which merely employs a 3D CNN encoder–decoder, adding the short-term or long-term temporal feature module will improve MAE and RMSE. By contrast, concurrently using both modules substantially reduced both types of errors. Thus, we can observe that both the modeling of short-term and long-term temporal features contribute to the performance improvement of the research algorithm. The simultaneous utilization of the two has significantly enhanced the algorithm's capacity for temporal feature modeling and conspicuously elevated the overall performance of the baseline architecture.

4. Conclusions

This paper proposes a method for extracting rPPG signals based on the baseline architecture of the 3D CNN encoder–decoder. Furthermore, a long-term and short-term temporal

self-attention module is introduced based on this baseline architecture for extracting and integrating short-term and long-term temporal features, enhancing the spatial-temporal feature modeling capability of the overall network architecture. Comparative experiments on video length and the cropping size of facial regions were conducted during the experimental stage to determine the appropriate data size. This study carried out experiments on the UBFC-rPPG and PURE datasets to validate the superiority of the proposed method. The experimental results demonstrated the state-of-the-art performance of the proposed method in this study.

Author Contributions: Software, Z.Z.; Data curation, J.L.; Writing—original draft, Z.Z.; Writing—review & editing, J.S., L.Z. and S.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The UBFC-rPPG can be found from https://sites.google.com/view/ybenezeth/ubfcrppg (accessed on 1 October 2024). The Pure can be found from https://www.tu-ilmenau.de/en/university/departments/department-of-computer-science-and-automation/profile/institutes-and-groups/institute-of-computer-and-systems-engineering/group-for-neuroinformatics-and-cognitive-robotics/data-sets-code/pulse-rate-detection-dataset-pure (accessed on 1 October 2024).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Horigome, H.J.P.C.; Surgery, C. Let's Recognize again the Usefulness of Electrocardiography in Cardiovascular Medicine. *J. Pediatr. Cardiol. Card. Surg.* **2018**, *34*, 1–2. [CrossRef]
2. Jong, G.-J.; Aripriharta; Horng, G.-J. The PPG Physiological Signal for Heart Rate Variability Analysis. *Wirel. Pers. Commun.* **2017**, *97*, 5229–5276. [CrossRef]
3. Poh, M.Z.; McDuff, D.J.; Picard, R.W. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Opt. Express* **2010**, *18*, 10762–10774. [CrossRef] [PubMed]
4. Guazzi, A.R.; Villarroel, M.; Jorge, J.; Daly, J.; Frise, M.C.; Robbins, P.A.; Tarassenko, L. Non-contact measurement of oxygen saturation with an RGB camera. *Biomed. Opt. Express* **2015**, *6*, 3320–3338. [CrossRef] [PubMed]
5. Chen, W.; McDuff, D. DeepPhys: Video-Based Physiological Measurement Using Convolutional Attention Networks. In Proceedings of the Computer Vision—ECCV 2018, Cham, Switzerland, 8–14 September 2018; pp. 356–373.
6. Yu, S.N.; Wang, C.S.; Chang, Y.P. Heart Rate Estimation From Remote Photoplethysmography Based on Light-Weight U-Net and Attention Modules. *IEEE Access* **2023**, *11*, 54058–54069. [CrossRef]
7. Massaroni, C.; Nicolò, A.; Sacchetti, M.; Schena, E. Contactless Methods For Measuring Respiratory Rate: A Review. *IEEE Sens. J.* **2021**, *21*, 12821–12839. [CrossRef]
8. Verkruysse, W.; Svaasand, L.O.; Nelson, J.S. Remote plethysmographic imaging using ambient light. *Opt. Express* **2008**, *16*, 21434–21445. [CrossRef] [PubMed]
9. Lewandowska, M.; Rumiński, J.; Kocejko, T.; Nowak, J. Measuring pulse rate with a webcam—A non-contact method for evaluating cardiac activity. In Proceedings of the 2011 Federated Conference on Computer Science and Information Systems (FedCSIS), Szczecin, Poland, 18–21 September 2011; pp. 405–410.
10. Wang, W.; Stuijk, S.; Haan, G.D. Exploiting Spatial Redundancy of Image Sensor for Motion Robust rPPG. *IEEE Trans. Biomed. Eng.* **2015**, *62*, 415–425. [CrossRef] [PubMed]
11. Wang, W.; Brinker, A.C.D.; Stuijk, S.; Haan, G.D. Algorithmic Principles of Remote PPG. *IEEE Trans. Biomed. Eng.* **2017**, *64*, 1479–1491. [CrossRef] [PubMed]
12. Spetlik, R.; Cech, J.; Franc, V.; Matas, J. Visual Heart Rate Estimation with Convolutional Neural Network. In Proceedings of the British Machine Vision Conference, Newcastle, UK, 3–6 September 2018.
13. Yu, Z.; Li, X.; Zhao, G. Remote Photoplethysmograph Signal Measurement from Facial Videos Using Spatio-Temporal Networks. In Proceedings of the British Machine Vision Conference, Cardiff, UK, 9–12 September 2019.
14. Liu, X.; Wei, W.; Kuang, H.; Ma, X. Heart Rate Measurement Based on 3D Central Difference Convolution with Attention Mechanism. *Sensors* **2022**, *22*, 688. [CrossRef] [PubMed]
15. Zhao, C.; Hu, M.; Ju, F.; Chen, Z.; Li, Y.; Feng, Y. Convolutional neural network with spatio-temporal-channel attention for remote heart rate estimation. *Vis. Comput.* **2023**, *39*, 4767–4785. [CrossRef]
16. Huang, B.; Lin, C.-L.; Chen, W.; Juang, C.-F.; Wu, X. A novel one-stage framework for visual pulse rate estimation using deep neural networks. *Biomed. Signal Process. Control* **2021**, *66*, 102387. [CrossRef]
17. Wang, Z.K.; Kao, Y.; Hsu, C.T. Vision-Based Heart Rate Estimation Via A Two-Stream CNN. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 3327–3331.

18. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I.J. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762. [CrossRef]
19. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929. [CrossRef]
20. Yu, Z.; Shen, Y.; Shi, J.; Zhao, H.; Torr, P.; Zhao, G. PhysFormer: Facial Video-based Physiological Measurement with Temporal Difference Transformer. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 4176–4186.
21. Liu, X.; Hill, B.; Jiang, Z.; Patel, S.; McDuff, D. EfficientPhys: Enabling Simple, Fast and Accurate Camera-Based Cardiac Measurement. In Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2–7 January 2023; pp. 4997–5006.
22. Yu, F.; Koltun, V.J. Multi-Scale Context Aggregation by Dilated Convolutions. *arXiv* **2015**, arXiv:1511.07122. [CrossRef]
23. Wang, X.; Girshick, R.; Gupta, A.; He, K.J. Non-local Neural Networks. *arXiv* **2017**, arXiv:1711.07971. [CrossRef]
24. Bobbia, S.; Macwan, R.; Benezeth, Y.; Mansouri, A.; Dubois, J. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognit. Lett.* **2019**, *124*, 82–90. [CrossRef]
25. Stricker, R.; Müller, S.; Gross, H.M. Non-contact video-based pulse rate measurement on a mobile service robot. In Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication, Edinburgh, UK, 25–29 August 2014; pp. 1056–1062.
26. Akima, H. A New Method of Interpolation and Smooth Curve Fitting Based on Local Procedures. *J. ACM* **1970**, *17*, 589–602. [CrossRef]
27. Liu, X.; Narayanswamy, G.; Paruchuri, A.; Zhang, X.; Tang, J.; Zhang, Y.; Sengupta, S.; Patel, S.; Wang, Y.; McDuff, D.J. rPPG-Toolbox: Deep Remote PPG Toolbox. *arXiv* **2022**, arXiv:2210.00716. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.